The 2001 NIST Evaluation Plan for Recognition of Conversational Speech over the Telephone

Version 1.1, 30-Oct-00

Introduction

The 2001 evaluation of conversational speech recognition over the telephone is part of an ongoing series of periodic evaluations conducted by NIST. These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of conversational speech recognition. To this end the evaluation is designed to be simple, to focus on core speech technology issues, to be fully supported, and to be accessible.

This year's evaluation is on conversational telephone data in English. This year for the first time part of the evaluation test set will consist of cellular data. Evaluations using conversational telephone data in Mandarin and Spanish may also be conducted if there is sufficient interest. The primary evaluation metric will be word error rate (character error rate for Mandarin).

The 2001 evaluation will be conducted in February and March. A follow-up workshop for participants in this evaluation will be held in May (immediately before ICASSP) to discuss research findings. The specific dates are listed in the Schedule given below.

As last year, the English evaluation will also have a non-competitive diagnostic component for which phone level results will be requested. Participation in this component of the evaluation will be optional, and details of it will be announced later in a separate plan document.

Participation in this evaluation is solicited for all sites that find the task worthy and the evaluation of interest. For more information, and to register a desire to participate in the evaluation, please contact Dr. Alvin Martin at NIST, alvin.martin@nist.gov. Please note that the commitment deadline for participation is February 1, 2001.

Separate mailing lists will be maintained of those interested in the Mandarin or the Spanish evaluation. Anyone with an interest in either, whether as a participant or otherwise, is asked to contact Dr. Martin at NIST.

Technical Objective

The Hub-5 evaluation focuses on the task of transcribing conversational speech into text. This task is posed in the context of conversational telephone speech in General American English (and perhaps in Mandarin and in Spanish). The evaluation is designed to foster research progress, with the goals of

1. exploring promising new ideas in the recognition of conversational speech,
2. developing advanced technology incorporating these ideas, and
3. measuring the performance of this technology.

The Task

The task is to transcribe conversational speech, which is presented as a set of conversations or parts of conversations collected over the telephone. The speech data is represented as a "4-wire" recording, that is, with two distinct sides,

one from each end of the telephone circuit. Each side is recorded and stored as a standard telephone codec signal (8 kHz sampling, 8-bit mu-law encoding).

The speech data is represented as a sequence of "utterances", where each utterance is a period of time when one speaker is speaking. Successive utterances may be speech segments of either the same speaker or of both conversation participants. The transcription task is to produce the correct transcription for each of the specified utterances. The beginning and ending times[1] of each of these utterances will be supplied as side information to the system under test. This information, stored in a single PEM file[2], will determine the test material.

Speech Data

Transcription Conventions

The American Heritage Dictionary (AHD)[3] will serve as the standard reference for word spellings in English. Words that don't occur in the AHD will be spelled using the most common accepted spelling. The official lexicons supplied by the Linguistic Data Consortium will define the standard spellings in Spanish and Mandarin.

Hesitation sounds, referred to as "non-lexemes", will be represented with a leading "%" character. Although these sounds are transcribed in a variety of ways due to highly variable phonetic quality, they are all considered to be functionally equivalent from a linguistic perspective.

Training Data

The following may be used for English language training:

- the entire SwitchBoard (i.e., Switchboard-1) Corpus as released,
- the entire Switchboard-2 Phase-1 Corpus, and
- all English conversations of the Call_Home Corpus, including those originally designated for training and those used as test data in previous evaluations.

These corpora are available from the LDC. Note that while the entire Switchboard-2 Phase-1 Corpus may be used, only the 20 conversations from it in the 1997 test set have been transcribed. Note also that no CallHome English conversations are included in this year's evaluation test set (described below). Additional English data may also be used for training, provided that the data are publicly available at the time of reporting results.

In Spanish and Mandarin the following may be used for training:

- all of the Call_Home conversations in each language that are designated as training data
- the 20 conversations in each language originally designated as development data
- the Call_Home conversations in each language used as evaluation data in 1995 and 1996 and, for Mandarin only, as evaluation data in 1997.

These corpora are available from the LDC. Additional Spanish and Mandarin data may also be used for training, provided that the data are publicly available at the time of reporting results.

The CASS Corpus of annotated Chinese speech will also be available to all participants for system development work (see The Evaluation below).

Development Data (the DevSet)

The English development data will consist of the following three distinct parts, corresponding to the three parts of the English EvalSet described below:

- **DevSet-1:** the 20 original Switchboard-1 conversations in the 1999 EvalSet (not included in the published corpus as originally released). Note that most of the speakers in these conversations also appear in the released Switchboard Corpus.

- **DevSet-2:** the 20 Switchboard-2 Phase-2 conversations in the 1998 EvalSet

- **DevSet-3:** a set of 20 cellular conversations from the Switchboard-2 Phase-4 Corpus

The first two of these three parts have been distributed to previous evaluation participants and are available from the LDC. The cellular development conversations will be available from NIST by December 1. Segment time marks (STM) and corresponding standard normal orthographic representation (SNOR) transcriptions for this data will be provided in standard STM[4] format.

In addition, subsets of each part of this DevSet will be defined for the purpose of facilitating exchange of research results between sites. These will consist of the first 30 seconds of speech (within the 5-minute evaluation excerpt) from each conversation side. (The elapsed time will be 30 seconds or more, in order to capture 30 seconds of actual speech. In some cases, the elapsed time may reach the limit of 5 minutes if a speaker mostly listens.) Segment time marks and corresponding transcriptions for these subsets will be provided by NIST in standard STM format

The 1997 Call_Home Spanish and 1999 Call_Home Mandarin EvalSets (20 conversations in each language) will serve as the DevSets for the respective languages. Segment time marks (STM) and corresponding SNOR transcriptions for these data will be provided in standard STM format

Evaluation Data (the EvalSets)

The English evaluation data will consist of the following three distinct parts:

- **EvalSet-1:** 20 conversations collected for the original Switchboard Corpus, but not included in the original release. Note that most of the speakers in these conversations also appear in the released Switchboard Corpus.

- **EvalSet-2:** 20 conversations from the Switchboard-2 Phase-3 Corpus.

- **EvalSet-3:** 20 cellular conversations from the new Switchboard-2 Phase-4 Corpus.

All 60 conversations, each of five minutes duration, will be provided on a single CDROM.

Speaker utterance segmentation information for all of the EvalSet will be supplied to guide the recognition system. This segmentation information will be supplied in NIST's PEM file format.

While the data on the CD, consisting of 60 whole conversations, will come from three different sources, the identity of the source is not to be provided to the system under test. The system must either recognize the speech irrespective of the source or must automatically determine the source from examination of the speech signal.

The Mandarin EvalSet will consist of 20 conversations from the Call_Home corpus and will be distributed on a separate evaluation CDROM. Likewise, the Spanish EvalSet will consist of 20 conversations from the Call_Home corpus and will be distributed on a separate evaluation CDROM. Speaker utterance segmentation information will be supplied to guide the recognition system. This segmentation information will be supplied in NIST's PEM file format.

The Evaluation

Each system will be evaluated by measuring that system's word error rate (WER). (For Mandarin the measure will be character error rate (CER), and all references to words in what follows should be taken as referring to characters.) Each system will also be evaluated in terms of its ability to predict recognition errors. System performance will be evaluated over an ensemble of conversations and parts of conversations. Performance will be separately evaluated for each of the three parts of the English evaluation set. The content of each of these parts is being chosen, to the

extent possible, to represent a statistical sampling of conditions of evaluation interests. These conditions will include sex, geographical distribution, and age.

For the English evaluation participating sites may choose to run their systems on all three parts of the EvalSet or on only one or two of these parts. Evaluation must be done on all twenty conversations of whichever part or parts are chosen. A single baseline system should be run on all parts for which the site is participating. Additional alternate systems are welcome, and these may include systems specifically designed or tuned for one part or parts of the data, such as the cellular data, and run only on that part.

As noted above, the speakers in the unreleased original Switchboard evaluation conversations (EvalSet-1) appear elsewhere in the published corpus, and thus in the training data. Results from last year's evaluation suggest that performance will not benefit greatly from this circumstance. The identities of these speakers will be made available before the evaluation. Thus sites may, if interested, create alternate systems which include retraining without these speakers.

It is hoped that the proposed Mandarin evaluation can build on the resources and work done on Mandarin recognition at the workshop at Johns Hopkins this past summer. The CASS Corpus of annotated Chinese spontaneous (but not telephone or conversational) speech will be made available to participants for system development work. This annotated corpus of spontaneous speech was developed in China to support research in this area.

The Reference Transcription

EvalSet-1 and EvalSet-2 are being transcribed at NIST in accordance with the guidelines in the document "General Instructions for SWITCHBORAD Transcriptions".[5] Transcription of much of the Switchboard-2 Phase-4 cellular corpus, of which EvalSet-3 is a part, is being done at the LDC, following somewhat older guidelines. NIST will transform the "turns" of these transcripts into the sometimes shorter "utterances" specified by these newer guidelines.

The reference transcriptions are intended to be as accurate as possible, but there will necessarily be some ambiguous cases and outright errors. In view of the existing high error rates of automatic recognizers on this type of data, it is not considered cost effective to generate multiple independent human transcriptions of the data or to have a formal adjudication procedure following the evaluation submissions.

The reference transcription for each utterance will be limited to a single sequence of words. This word sequence will represent the transcriber's best judgment of what the speaker said.

Word fragments will be represented by an initial part of a word with a hyphen at the end. Correct recognition will consist of either ignoring the fragment, or producing a word of which the fragment is an initial part.

The reference transcription will contain no hyphenated words. Each hyphenated word will be separated into its separate constituent words

The WER (CER for Mandarin) Metric

Word error rate is defined as the sum of the number of words in error divided by the number of words in the reference transcription. The words in error are of three types, namely ***substitution*** errors, ***deletion*** errors, and ***insertion*** errors. Identification of these errors results from the process of aligning the words in the reference transcription with the words in the system output transcription. This alignment is performed using NIST's SCLITE software package[6].

- A substitution error results when the spellings of the reference word and the corresponding system output word differ.

- A deletion error results when the reference word has no corresponding system output word.

- An insertion error results when a system output word has no corresponding reference word.

Scoring will be performed by aligning the system output transcription with the reference transcription and then computing the word error rate. Alignment will be performed independently for each utterance. The system output transcription will be processed to match the form of the reference transcription. Hyphenated words will be separated into their separate constituent words.

A few variant spellings of the same word exist in the English transcriptions. These words, with or without hyphens, will be mapped onto a single preferred word spelling without hyphens. The set of all such mappings is:

Input	Output
mhm	uhhuh
mmhm	uhhuh
mm-hm	uhhuh
mm-huh	uhhuh
huh-uh	uhuh

For scoring purposes, all hesitation sounds will be considered to be equivalent. Thus all reference transcription words beginning with "%", the hesitation sound flag, along with the conventional set of hesitation sounds, will be mapped to "%hesitation".

The system output transcriptions should use any of the hesitation sounds (without "%") when a hesitation is hypothesized. The set of English hesitation sounds for the current evaluation is defined to be:

"uh", "um", "eh", "mm", "hm", "ah", "huh", "ha", "er", "oof", "hee", "ach", "eee" and "ew".

The sets of hesitation sounds for Spanish and Mandarin may be found at 1997 evaluation website:

ftp://jaguar.ncsl.nist.gov/evaluations/hub5ne/sept97/datafiles/current_datafiles.htm.

As noted previously, for Mandarin character error rate will be used in place of word error rate. Furthermore, confidence scores, as discussed below, will be applied at the character level. If the system output gives confidences only at the word level, the word level values will be automatically imputed to characters making up the word.

The Confidence Measure

Along with each word output by a system, a confidence score is also required. This confidence score is the system's estimate of the probability (in the range [0,1]) that the word (or character for Mandarin) is correct. While this might be merely a constant probability, independent of the input, certain applications and operating conditions may derive significant benefit from a more informative estimate that is sensitive to the input signal. This benefit will be evaluated by computing a normalized cross entropy (NCE) measure consisting of the mutual information (cross entropy) between the correctness of the system's output word and the confidence score output for it, normalized by maximum cross entropy:

$$NCE = \left\{ H_{\max} + \sum_{correct\ w} \log_2\left(\hat{p}(w)\right) + \sum_{incorrect\ w} \log_2\left((1-\hat{p}(w)\right) \right\} \Big/ H_{\max}$$

where,

$$H_{\max} = -n\log_2(p_c) - (N-n)\log_2(1-p_c)$$
$$n = \textit{the number of correct HYP words}$$
$$N = \textit{the total number of HYP words}$$
$$\hat{p} = \textit{the confidence measure output, as a function of output word}$$
$$p_c = \textit{the average probability that an output word is correct} = n/N$$

In addition, as in the past, NIST will use the likelihood scores along with the hypothesized words to create a DET (Detection Error Tradeoff) type curve for each set of results submitted. This will be a plot of

$$(\#\text{substitutions} + \#\text{insertions}) / \#(\text{hypothesized words}) \text{ vs. } (\#\text{substitutions} + \#\text{deletions}) / \#(\text{reference words})$$

as the likelihood score is varied and used as a threshold for determining whether hypothesized words should be included.

Submission of Results

All results must be submitted by 7:00 PM EST on 12 March 2001. Sites should submit results using the following steps:

1. system output file creation,
2. directory structure creation,
3. system documentation, *including execution times,* and system output inclusion
4. transmission protocol to NIST.

Step 1: System output file creation

The time-marked hypothesis tokens for each test will be placed in a single file, called "<TEST_SET>.ctm", where <TEST_SET> is the base name of the associated PEM file. The CTM (Conversation Time-Mark) file format is a concatenation of time marks for each hypothesized token in each side of a conversation. Each hypothesized token must have a conversation id, channel identifier [A | B], start time, duration, case-insensitive text, and a confidence score. The start time must be in seconds and relative to the beginning of the waveform file. The conversation id's for this evaluation will be of the form:

CONV_ID::= <SWB_TYPE_ID > | <SWB2_ID> | <SWB_CELLULAR_ID>

where,

SWITCHBOARD_TYPE_ID ::= swDDDD (where DDDD is a four digit conversation code)

SWB2_ID ::= sw_DDDDD (where DDDDD is a five digit conversation code)
SWB_CELLULAR_ID ::= sw_4DDDD (where DDDD is a four digit conversation code)

For the Mandarin evaluation, sites that choose to supply confidence scores at the character level must create a separate CTM record for each character. Otherwise, confidence scores for multi-character words will be imputed to all characters.

The file must be sorted by the contents of the first three columns: the first and the second in ASCII order,

the third in numeric order. The UNIX sort command: "sort +0 -1 +1 -2 +2nb -3" will sort the words into appropriate order.

Lines beginning with ';;' are considered comments and are ignored. Blank lines are also ignored.

Included below is an example:

```
;;
;; Comments follow ';;'
;;
;; The Blank lines are ignored

;;
en_7654 A 11.34 0.2 YES -6.763
en_7654 A 12.00 0.34 YOU -12.384530
en_7654 A 13.30 0.5 CAN 2.806418
en_7654 A 17.50 0.2 AS 0.537922
:
en_7654 B 1.34 0.2 I -6.763
en_7654 B 2.00 0.34 CAN -12.384530
en_7654 B 3.40 0.5 ADD 2.806418
en_7654 B 7.00 0.2 AS 0.537922
:
```

Step 2: Directory Structure Creation

Create a directory identifying your site ('SITE') which will serve as the root directory for all your submissions. Examples:

- bbn
- dragon
- ibm
- sri
- …

You should place all of your recognition test results in this directory. When scored results are sent back to you and subsequently published, this directory name will be used to identify your organization.

For each test system, create a sub-directory under your 'SITE' directory identifying the system's name or key attribute. The sub-directory name is to consist of a free-form system identification string 'SYSID' chosen by you. Place all files pertaining to the tests run using a particular system in the same SYSID directory.

The following is the BNF directory structure format for Hub-5 hypothesis recognition results:

<SITE>/<SYSID>/<FILES>

where

SITE ::= bbn | dragon | ibm | sri | . . .
SYSID ::= (short system description ID, preferably <= 8 characters)
FILES ::=

sys-desc.txt :: system description, including reference to paper if applicable

<TEST_SET>.ctm :: file containing time-marked hypothesis word.

where

TEST_SET ::= base name of the corresponding PEM file.

Step 3: System Documentation, *including execution times,* and System Output Inclusion

For each test you run and for each system evaluated, a brief description of the system (the algorithms) used to produce the results must be provided along with the results. (It is permissible for a single site to submit multiple systems for evaluation. In this case, however, the submitting site must identify one system as the "primary" system prior to performing the evaluation.)

The format for the system description is as follows:

SITE/SYSTEM NAME
TEST DESIGNATION

1. Primary Test System Description:
2. Acoustic Training:
3. Grammar Training:
4. Recognition Lexicon Description:
5. Differences for each Contrastive Test: *(if any contrastive test were run.)*
6. New Conditions for This Evaluation:
7. Execution Time:
   Sites must report the CPU execution time that was required to process the test data, as if the test were run on a single CPU. Sites must also describe the CPU and the amount of memory used.
8. References:

Step 4: Test Results Submission Protocol

Once you have structured all of your recognition results according to the above format, you can then submit them to NIST. Due to international e-mail file size restrictions, test sites are permitted to submit results to NIST using either email or anonymous ftp. Continental US sites may use either method, but international sites must use the 'ftp' method. The following instructions assume that you are using the UNIX operating system. If you do not have access to UNIX utilities or ftp, please contact NIST to make alternate arrangements.

E-mail method:

First change directory to the directory immediately above the <SITE> directory. Next, type the following:

```
tar -cvf - ./<SITE> | compress | uuencode <SITE>-<SUBM_ID>.tar.Z | \
mail -s "March 2001 Hub-5 test results <SITE>-<SUBM_ID>" \
alvin.martin@nist.gov
```

where

<SITE> is the name of the directory created in Step 2 to identify your site.

<SUBM_ID> is the submission number (e.g. your first submission would be numbered '1', your second, '2', etc.)

Ftp method:

First change directory to the directory immediately above the <SITE> directory. Next, type the

following command.

tar -cvf - ./<SITE> | compress | <SITE>-<SUBM_ID>.tar.Z

where

<SITE> is the name of the directory created in Step 2 to identify your site.
<SUBM_ID> The submission number (e.g. your first submission would be numbered '1', your second, '2', etc.)

This command creates a single file containing all of your results. Next, ftp to jaguar.ncsl.nist.gov giving the username 'anonymous' and your e-mail address as the password. After you are logged in, issue the following set of commands, (the prompt will be 'ftp'):

- ftp> cd /incoming/
- ftp> binary
- ftp> put <SITE>-<SUBM_ID>.tar.Z
- ftp> quit

You've now submitted your recognition results to NIST. The last thing you need to do is send an e-mail message to Alvin Martin at 'alvin.martin@nist.gov' notifying NIST of your submission. Please include the name of your submission file in the message.

Note:

If you choose to submit your results in multiple shipments, please submit ONLY one set of results for a given test system/condition unless you've made other arrangements with NIST.

Otherwise, NIST will programmatically ignore duplicate files.

Schedule

Cellular DevSet Release	1 December 2000
Commitment Deadline	1 February 2001
EvalSet Release	12 February 2001
Results Due at NIST	12 March 2001 at 7:00 PM EST (midnight GMT)
Results Release	26 March 2001
Workshop	3-4 May 2001

End Notes

[1] These utterance time marks will be specified in seconds (to the nearest millisecond) and will completely encompass the utterance. Thus successive utterances will overlap when the speakers talk over each other.

[2] The PEM ("partitioned evaluation map") file format is given in the SCLITE documentation available through NIST's web page (http://www.nist.gov/itl/div894/894.01/software.htm). Each record contains 5 fields: <filename>, <channel ("A" or "B")>, <speaker ("unknown")>, <begin time> and <end time>.

[3] The American Heritage Dictionary of the English Language, Book and CD ROM. Published October 1994 by Houghton Mifflin. ISBN 0395711460.

[4] STM stands for "segment time marked". The STM file identifies time intervals along with the transcription for those intervals. At the time this document was prepared, the STM file format is documented in NIST's SCLITE scoring software distribution available via NIST's web page (http://www.nist.gov/itl/div894/894.01/software.htm).

[5] http://www.isip.msstate.edu/projects/switchboard/doc/transcription_guidelines/

[6] SCLITE software is available via NIST's web page (http://www.nist.gov/itl/div894/894.01/software.htm).